MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH

Konrad-Zuse-Strasse 1 · D-18057 Rostock · Germany · Tel +49 (0) 3 81 20 81 - 0 · Fax +49 (0) 3 81 20 81 - 202 · www.demogr.mpg.de

MPIDR Working Paper WP 2023-012 I March 2023
https://doi.org/10.4054/MPIDR-WP-2023-012

Online Social Integration of Migrants: Evidence from Twitter

Jisu Kim I kim@demogr.mpg.de
Soazic Elise Wang Sonne
Kiran Garimella
André Grow I grow@demogr.mpg.de
Ingmar Weber
Emilio Zagheni I office-zagheni@demogr.mpg.de

Online Social Integration of Migrants: Evidence from Twitter

Jisu Kim[1,*], Soazic Elise Wang Sonne[2], Kiran Garimella[3], André Grow[1], Ingmar Weber[4], Emilio Zagheni[1]
[1]Max Planck Institute for Demographic Research, Germany, kim@demogr.mpg.de
[2]World Bank, USA
[3]Rutgers university, USA
[4]Saarland University, Germany

* Corresponding author: Max Planck Institute for Demographic Research, Germany.
Email: kim@demogr.mpg.de

Abstract

As online social activities have become increasingly important for people's lives and well-being, understanding how migrants integrate into online spaces is crucial for providing a more complete picture of integration processes. We curate a high-quality data set to quantify patterns of new online social connections among immigrants in the United States. Specifically, we focus on Twitter, and leverage the unique features of these data, in combination with a propensity score matching technique, to isolate the effects of migration events on social network formation. The results indicate that migration events led to an expansion of migrants' networks of friends on Twitter in the destination country, relative to those of users who had similar characteristics, but who did not move. We found that male migrants between 19 and 29 years old who actively posted more tweets in English after migration also tended to have more local friends after migration compared to other demographic group, which indicates that migrants' demographic characteristics and language skills can affect their level of integration. We also observed that the percentage of migrants' friends who were from their country of origin decreased in the first few years after migration, and increased again in later years. Finally, unlike for migrants' friends networks, which were under their control, we did not find any evidence that migration events expanded migrants' networks of followers in the destination country. While following users on Twitter in theory is not a geographically constrained process, our work shows that offline (re)location plays a significant role in the formation of online networks.

Keywords— Big data, Gender, International migration, Social integration, Twitter

1. Introduction

The integration of migrants is a multidimensional process that involves various aspects of migrants' lives and of host societies, including the economic, educational, cultural, civil, and other characteristics of both the host and the sending countries. While understanding integration processes is challenging given their complexity, it is nonetheless important, as integration enables migrants to transition more smoothly to their new home. In addition, from the point of view of the host country, the successful integration of migrants creates a more harmonious society, and reduces inequality, social conflicts, and social polarisation. Measures of integration often rely on employment, education, and health statistics or related information mostly based on surveys. In some contexts, field experiments have been carried out to study the social integration of migrants (see, for instance, Kearns and Whitley [2015], Yoon [2021], Wessendorf and Phillimore [2019], Glorius et al. [2020]). However, these established data sources have limitations. For example, the sample size of these surveys is often small. Moreover, conducting a survey is costly, and the limited scope of such surveys may reduce the diversity of immigrants' countries of origin that can be studied. In recent years, a growing number of studies on the integration of migrants have used social media data, such as data from Twitter [Sîrbu et al., 2020, Kim et al., 2021b], to complement analyses based on survey data. Researchers have shown that because of the detailed information on social networks that can be obtained from social media data – including information on users' friends and followers, and on the opinions and information the users shared – analysing these data can provide new insights into integration processes. Most importantly, as online social activities have become important to people's lives and well-being, understanding the online space has become crucial for researchers seeking to paint a more complete picture of integration processes, and to understand the social determinants of inequalities. For many users of an online social platform such as Twitter, who tend to be younger than the general population, the value of their online community has increased, and may even be comparable to that of their offline community [Lehdonvirta and Räsänen, 2011].

Among the various aspects of migrant integration, we focus here on social integration. More specifically, we are interested in investigating whether recent immigrants have integrated socially into online spaces through Twitter. We aim to answer the following questions in this study: *After migration, does the number of friends/followers an immigrant has in an online space who are based in the destination country increase? Do immigrants lose their connections with their friends/followers in their origin country? How do the characteristics of migrants whose number of online friends living in the destination country increases differ from those of migrants who do not experience such an increase?*

It is important to study the social integration of migrants, as it is one of the main dimensions of migrants' experiences, and it can facilitate various aspects of their integration process. Migrants' social networks often become their main sources of information, ranging from trivial information like "where to go for a nice dinner in the area"; to more valuable information, such as "my company is looking for an employee". Thus, the barriers to information on the local community migrants often encounter may be removed through the sharing of knowledge by local friends [Rauch, 2001, Rauch and Trindade, 2002]. At the same time, understanding migrants' social connections to people in their country of origin is also crucial, as it can shed light on immigrants' economic activities, such as the sending of remittances; and on the diffusion of culture and global trade at the macro level [Docquier and Rapoport, 2012, Rauch, 2001, Bahar et al., 2020]. In this study, we focus on immigrants residing in the United States. Studying the social integration of immigrants in the U.S. is particularly interesting, as it is known for being "the most diverse country in the world"[1].

[1] https://www.pewresearch.org/fact-tank/2020/08/20/key-findings-about-u-s-immigrants/

Our methodology includes ensuring a careful design of the data collection process, identifying migrants using geo-tagged tweets, and, finally, determining the dates when migrants made social connections. Using Twitter data that we obtained from archive.org, we start by identifying migrants on Twitter along with their likely date of migration, based on our definition of a migrant: namely, "a user who tweeted at least one geo-tweet per month in one country (home country) for 12 consecutive months and one geo-tweet per month for 12 consecutive months in another country (destination country)". This approach allows us to identify recent immigrants. The day a migrant started to tweet in their destination country is identified as the person's date of migration. Second, using the method developed by Meeder et al. [2011], we infer the dates when social links were established to study how the composition of the migrant's social network changed *before* and *after* migration. In this work, we prioritise having reliable and high-quality information on users over the size of the data. Hence, we match migrants with non-migrants with similar characteristics on Twitter using the 1:1 propensity score matching technique [Stuart et al., 2011]. This matching technique allows us to also provide a *placebo* date of migration to non-migrants using the same date of actual migration for the matched migrants. This enables us to compare the friendship patterns of these two groups, and to identify, in a quasi-causal sense, the effects of the migration event on the expansion of migrants' social links on Twitter. With the enriched data that we curated, we then examine the online social integration patterns of both migrants and non-migrants by looking at how many new U.S.- based and origin-based friends/followers were added to the migrants' networks after their likely date of migration.

The rest of this paper is organised as follows. In the next section, we provide background information, drawing from the sociological literature that deals with social integration, and from the emerging literature that leverages digital trace data. In the third section, we describe the data and methods we use. In several subsections, we provide details on our data, and explain how we identified migrants, validated the data, detected new social connections, and matched migrants and non-migrants. In the fourth section, we present our results. We offer our concluding remarks in Section 5.

2. Related works

The topic of migrant integration has long attracted the interest of social scientists across a number of disciplines, including sociology, economics, anthropology, and psychology, as the process of integration involves a broad spectrum of individual and societal characteristics. One of the concepts related to integration that sociologists have studied extensively is that of "social capital" [Bourdieu, 1986, Portes, 1998]. Bourdieu and Wacquant defined this concept as follows: "Social capital is the sum of the resources, actual or virtual, that accrue to an individual or a group by virtue of possessing a durable network of more or less institutionalised relationships of mutual acquaintance and recognition" [Wacquant and Bourdieu, 1992]. The core idea is that social connections are valuable assets that enable members of a society to forge mutually beneficial relationships. Social capital facilitates the exchange of information and knowledge, and increases economic, social, and cultural opportunities for individuals by, for example, pro-

viding them with job security, and social ties. Barriers to obtaining relevant information that immigrants would otherwise face are reduced through having ties in the host society [Putnam, 1995, Schiff, 1992, Uslaner, 2003].
The existing body of literature on the social integration of migrants has examined its mechanisms from different societal points of view, while relying mainly on surveys, interviews, or questionnaires. These surveys usually include elements specific to the contexts where they were conducted. For instance, to examine the determinants of social integration in the context of internal migration in China, Chen and Wang [2015] asked respondents "whether they actively participated in local community activities, whether they had adapted to local social norms and customs and whether they are socially connected with the local community". Based on the responses to these questions, the authors found that education and receiving higher income were key factors for social integration, as they provided migrants with the time and the resources they needed to integrate into local society. They also reported that migrants from geographically close regions were more likely to feel socially integrated. Similarly, Kearns and Whitley [2015] examined different aspects of social integration from the point of view of trust, reliance, safety, and sense of community in the UK. They then analysed the associations between these social integration factors and the effects of time, place, and migrant type; and of functional factors, such as educational qualifications, language proficiency, and employment status. The respondents were asked, for example, to what extent they felt that their "neighbourhood is a place where neighbours look out for each other"; and to what extent they agreed with the following statements: "It is likely that someone would intervene if a group of youths were harassing someone in the local area"; and "someone who lost a purse or wallet around the area would be likely to have it returned without having anything missing". The findings of this study are similar to those of Chen and Wang [2015]: i.e., that all functional factors were positively associated with social integration. In particular, they found that higher educational levels were positively associated with greater use of local amenities, and that being employed allowed migrants to access the resources they needed to interact with other members of the community. As the authors were looking at migrants in the UK, they also found that the migrants' language proficiency was positively associated with their level of trust, as it facilitated their communication with their neighbours. Similarly, Yoon [2021] examined the current integration status of North Korean migrants in South Korea. Among other indications of integration, the study investigated the social connections of these migrants, which the author defined using three criteria: number of South Korean friends ("social bridges"), number of North Korean friends who could help in an emergency ("social bonds"), and access to support services from the government ("social links"). The results indicated that North Koreans had a strong sense of belonging and trust. Indeed, the author observed that "North Korean migrants had strong belief that they could reach the same status as South Koreans", and were often identified as South Korean citizens.

These are examples of studies that have provided detailed information on migrants' experiences with social integration. However, there are also drawbacks to relying on surveys or interviews only. For instance, most studies of integration have been conducted for specific settings (e.g., in their neighbourhoods [Chen and Wang, 2015], migrants in the UK [Kearns and Whitley, 2015, Wessendorf and Phillimore, 2019], and North Koreans in South Korea [Yoon, 2021]). This means that the geographical coverage of existing studies is often limited, as they tend to be concentrated on country-specific contexts. Moreover, in a survey, respondents may feel compelled to choose an answer even if their views do not fit any of the choices that they are given, or to agree with given statements (e.g., acquiescence bias). Additionally, information on respondents' social networks can be difficult to obtain. In many cases, only a small part of respondents' social groups are captured, or their networks can differ depending on when the interview was conducted. For example, if a migrant had met a group of friends recently, they may remember them the most clearly, which could influence their survey responses.

In recent years, social media data such as data from Facebook and Twitter have been leveraged to study a number of dimensions of migrant integration [Dubois et al., 2018, Stewart et al., 2019, Chi et al., 2019, Kim et al., 2021a,b, Mazzoli et al., 2020]. The use of social media data can overcome some of the limitations of traditional data sources, and can complement them. For instance, Dubois et al. [2018], Stewart et al. [2019] both extracted data from the Facebook advertising platform to study cultural assimilation. In both studies, the authors were able to obtain aggregated information on individual users' interests in specific areas through Facebook data. Dubois et al. [2018] quantified the level of assimilation of migrants in Germany by introducing a score that provides a proxy for the migrants' assimilation to the local population's interests. Using these data, the authors were able to list the detailed interests of Germans in various domains, including music, film, city, companies, and so on. They found that European migrants in Germany had a higher assimilation score than Arabic-speaking migrants and Turkish speakers. Going deeper into the data, they also found that among Arabic-speaking migrants, men, university graduates, and individuals aged 18-24 had higher assimilation scores. Focusing on Mexican immigrants of both the first and the second generation in the U.S., Stewart et al. [2019] looked at changes in their musical tastes as a specific case study of cultural assimilation. Using assimilation score metrics similar to those applied in Dubois et al. [2018], they examined the extent to which the migrants' musical tastes had converged to those of host population as evidence of their cultural assimilation. They were able to distinguish between the first and the second generation of migrants through the option of targeting specific users on the Facebook advertising platform. In general, they found that Mexican migrants of both the first and the second generation showed a high degree of assimilation to the musical tastes of Anglos and African Americans. They also found that younger migrants and the first generation of Mexican migrants had higher assimilation scores. However, in contrast to previous findings, they found that women assimilated more than men when it came to musical tastes. In addition, although the authors concentrated on the case study of Mexican immigrants, they also showed that their methodology could be used to study the wider immigrant population in the U.S. Also using Facebook data, Chi et al. [2019], studied the relationship between international social ties and human mobility through social network analysis. In this study, the authors had access to the social networks of individual users, which were not limited to one geographic setting, but could be worldwide. The results showed that long-term international migrants created 83% of the international ties.

Cultural integration has also been studied using Twitter data by Kim et al. [2021b]. They looked at the host country-specific hashtags used by migrants as a proxy for *destination attachment* level, and at the origin country-specific hashtags used by migrants as a proxy for *origin attachment* level, while also taking into consideration the preservation of links to the migrants' country of origin. With *destination attachment* and *origin attachment* indexes, they summarised integration patterns into four categories: marginalisation, separation, assimilation, and integration. Their results highlighted several factors that were positively related to different integration patterns. In particular, they observed that language proximity and the distance between the origin and the destination country played important roles in how the immigrants maintained their links to their origin country, and in how they integrated into the destination society. In addition, the authors were able to observe the behaviours of different migrants in various countries, including immigrants in the United States and Great Britain, and Italian emigrants in various destination countries. These kinds of patterns are difficult to observe using traditional data sources. Mazzoli et al. [2020] employed Twitter data to examine the spatial integration of migrants. In their study, the authors showed how Twitter data can be used to observe the places most commonly visited by Venezuelan migrants based on their tweet activity after working hours. Drawing on these observations, they built a segregation index by directly comparing the population distribution of migrants to natives, and the spatial segregation of Venezuelan migrants in Bogotà, Lima, and São Paulo. The results of their segregation index also showed that Venezuelan migrants were spatially segregated from the locals.

In this paper, we further advance the literature that uses Twitter data to study the integration of migrants. In contrast to the above mentioned studies, we focus on a different dimension of migrant integration (i.e., social integration) that has not been previously explored in the literature. Thanks to the availability of Twitter data, we are able to obtain information on the social networks of individual users. These data are different from publicly available Facebook data, which contain no information on the users' social networks. Our analyses build on the work of Chi et al. [2019], as we are also interested in the composition of the social networks of immigrants. Importantly, however, we have collected and harmonised information about the temporal order of the social links based on when they were created. Using these data, we have implemented various methodologies to build a high- quality data set that enables us to quantify how many new social connections migrants added on Twitter before and after migration. In addition, we further advance the literature by inferring whether the migrants' newly added social connections were or were not local friends/followers based on information provided on their profiles. These results allow us to determine to what degree migrants were socially integrated (on the Twitter platform) into the destination country. Moreover, in contrast to previous studies, we are interested in exploring the social integration patterns of migrants in an online setting, which is an important perspective given that the amount of time people spend on online social platforms has increased dramatically in the last decade.

3. Data and method

In order to study the online social integration of migrants, we focus on one specific platform: Twitter. We chose Twitter because the data available for this platform enable us to consider changes in the structure of the users' social networks *before* and *after* migration. The methodology includes: (i) the careful design of the data collection approach and of data pre-processing; (ii) the high-quality identification of migrants using geo-tagged tweets; and (iii) a solid statistical framework for analysing pre- and post-changes with respect to a properly selected control group.

3.1 Data

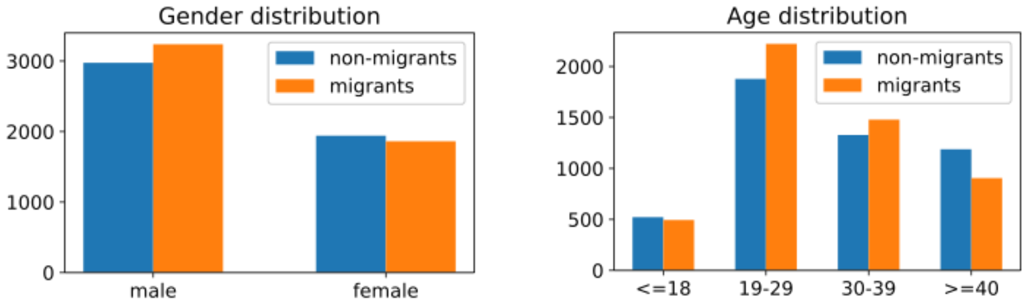


Figure 1: Distribution of demographic characteristics of users by migration status.

As a first step, we retrieved Twitter data from a large and publicly available archive of tweets (https://archive.org/details/twitterstream) that includes a random sample of 1% of all tweets.[2] This database of Twitter tweets includes, among other information, the text of each tweet, the uniquely identifying user IDs, the geo-location – if enabled by the user – of where the tweet was sent from, the language of the tweet, and the date and time when the tweet was created. As it is most crucial for our analysis to obtain location information, we filtered out tweets from 2017-2018 that were not geo-tagged. At this stage, the database included about one million users who had geo-tagged tweets from the time period we are interested in. To obtain the timeline of the geo-tagged tweets of these one million users, we collected their most recent 3200 tweets, their profile and social network information, as well as both their "friends" (Twitter users that a specific user follows) and "followers" (Twitter users that follow a specific user) using the Twitter API[3] . The time series of geo-tagged tweets of all of the users considered were important information for us, as they helped us to determine whether a user was or was not a migrant.

In order to augment the data with additional information, we used the deep learning algorithm[4] developed by Wang et al. [2019] to estimate the demographic characteristics of users (i.e., age and gender where the gender classification provides a binary category; and age in four broad categories: ≤ 18, (18,30), [30, 40), and [40, 99). Note that age was grouped into four categories following the age ranges used in censuses and surveys, but also because producing estimates at a higher level of granularity would result in extremely high uncertainty, even with the help of human coders [Wang et al., 2019]. This algorithm used as input the user's profile photo, biography, screen names,

[2] https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/api-reference/get- tweets-sample-stream

[3] https://developer.twitter.com/en/docs/twitter-api/v1/accounts-and-users/follow-search-get- users/overview

[4] https://github.com/euagendas/m3inference

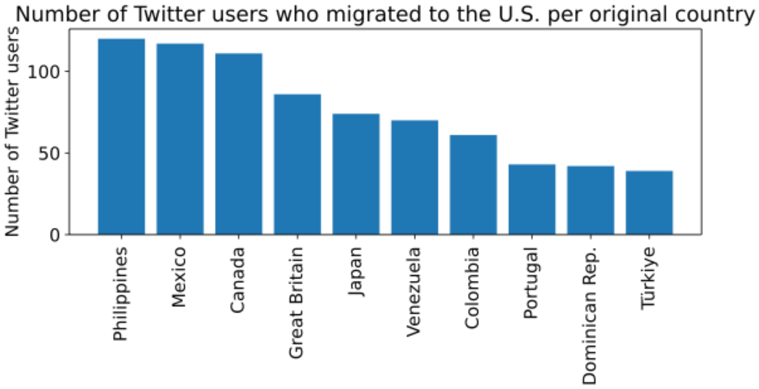


Figure 2: Top 10 countries of origin of the migrants in the United States in the analysed Twitter data set.

and names from the Twitter profile. To be more specific, the authors combined three models (image model, text model, and multimodal model) to process different types of information needed to estimate demographic characteristics. First, the image model was used to process the profile image using DenseNet [Huang et al., 2017]. Second, the text model was used to process the text inputs from the biography and the username using the Long-Short Term Memory (LSTM) architecture [Hochreiter and Schmidhuber, 1997]. Finally, the multimodal model was used to combine the first two models "into a new model with a modality drop-out layer" that allowed them to better regularise the model, including in cases in which not all of the input sources were available. To put it simply, the model picked up hints from the biographical information of users based on words or images that signal gender and age, such as "mother of two wonderful kids" or "26 y/o dude". The authors reported that when this algorithm's performance score on Twitter data was compared with other state-of-the- art algorithms, such as the Microsoft classifier[5] and Face++[6], it outperformed in all three categories. To be more specific, the gender classification obtained a Macro-F1 score of 0.92, 0.9 for organisation status classification, and 0.43 for age classification. Another advantage of this algorithm is that it also takes different languages into account, which allows us to obtain the demographic information of users who are from various countries. As Figure 1 shows, males made up the majority of our data set of Twitter users, regardless of migration status. Looking at the age distribution, we see that a sizeable proportion of our population of users were between 19 and 29 years old, regardless of migration status. We also observe that more migrants were concentrated in the central age groups, while more non-migrants were concentrated in the 40+ age group.

3.2 Identifying migrants

[5] https://azure.microsoft.com/en-us/services/cognitive-services/face/#overview

[6] https://www.faceplusplus.com

Provided with the timeline of geo-tagged tweets, we identify users as migrants if these users “have tweeted at least one geo-tagged tweet per month in one country (home country) for 12 consecutive months and one geo-tagged tweet per month for 12 consecutive months in another (destination) country”. In other words, we determine whether a user moved to a different country within the time frame of interest based on the tweet timeline of that user. For example, if a user who tweeted at least one geo-tagged tweet per month for 12 consecutive months from Korea was tweeting from the United States in the later period for 12 consecutive months, then we would identify this user as a migrant from Korea to the United States. Our definition of migrants is adapted from the United Nations, which defines a migrant as “a person who moves to a country other than that of his or her usual place of residence for a period of at least a year”. [7] Following this definition, we look at users’ geo-tagged tweets over a period of 12 consecutive months to ensure that we are not observing short-term visitors, such as students or tourists. We note also that the definition of migrants employed here focuses on users who moved to the U.S. during the time period we are interested in. Thus, users who migrated in the earlier period and did not change their location thereafter would not be captured by this definition. During this process, we also identify the likely date of migration, which is the date when the user started tweeting from the destination country.

By applying this approach, we were able to identify about 5000 migrants coming from 133 different countries who migrated to 154 countries. Of these destination countries, the United States attracted the largest number of migrants (1300), followed by Great Britain (440) and Brazil (321). Thus, we identified 1300 migrants to the United States, the majority of whom came from the Philippines, followed by Mexico, Canada, and the United Kingdom, as shown in Figure 2.

In parallel, we randomly sampled about 5000 non-migrants from the initial data (one million users), excluding the 5000 identified migrants. These non-migrants were selected to serve as a control group in the later stages of the analysis. Here, we define non-migrants as the users who tweeted in one country only during the same window of observation as the migrants, irrespective of whether they were or were not migrants. However, in this process, we also label the users who tweeted from several countries but did not stay longer than two months as non-migrants, based on the assumption that they were making short trips, and had not migrated. Similarly, long-term migrants who migrated before their move can be inferred from their Twitter timeline were classified as non-migrants. Implicitly, this means that our analytical setup assumes that the characteristics of long-term migrants were similar to those of U.S. residents. Like the migrants, most of the non-migrants in our data came from the United States, Brazil, Indonesia, and Great Britain.

3.3 Data validation

[7] Recommendations on Statistics of International Migration, Revision 1, Statistical Papers, Series M, No. 58, United Nations, New York, 1998, Glossary.

To validate the quality of both the inferred user demographics (age and gender), as well as the inferred date of migration, we manually checked 620 individual users in our data by looking at their public profiles on Twitter. Here, we checked for clues in their name, profile photo, self-declared bio, and tweets related to their demographics. Furthermore, we looked for any signs that users had migrated, such as any mention of a “new home”, changes in the geo-tags of their tweets, or photos of what seems to be a new home location. Note that the number of individual users for whom we manually checked the age category, mostly because their profile photo was missing, was lower (555). With this information, we computed an F1 score for each identified category to measure the performance of our estimation and classification. The F1 score measures the performance of our estimation, comparing our classification with true data (i.e., the 620 individual users we manually checked). It provides a score between zero and one, with one signifying perfect precision and recall. The precision score tells us, for instance, how many males among all the males we identified were indeed male; while the recall score tells us, for instance, how many males among all the males we identified were correctly predicted. In the following, we look at the weighted average F1 score, which considers the number of samples from each class.

Starting with the demographic characteristics of the users in the table, we obtained a weighted average F1 score of 0.92 for our gender categories, demonstrating the validity of our gender estimation procedure (see Table 1). Similarly, we obtained a weighted average F1 score of 0.85 for our age categories, which generally validates our age estimation procedure (see Table 2). The F1 score for age was slightly lower than the F1 score for gender due to the relatively low precision score for the population under age 18. Lastly, we obtained an accuracy level of 62% for the correct identification of the migration date of the users in our data. Of the remaining 38%, about 52% were users for whom we lacked any clear information about whether they truly were or were not migrants, while rest were users who either were correctly identified as migrants, but for whom the date of the migration was incorrectly estimated, or were short-term visitors (e.g., on vacation). It should be pointed out that despite the challenges involved in identifying migrants using Twitter data, as highlighted by Armstrong et al. [2021], we were able to accurately identify a considerable proportion of the migrants in our data.

	precision	recall	f1-score	support
Female	0.93	0.85	0.89	229
Male	0.92	0.96	0.94	391
accuracy	0.92	0.92	0.92	0.92
macro avg	0.93	0.91	0.92	620
weighted avg	0.92	0.92	0.92	620

Table 1: Average precision, recall and F1 scores for gender comparing our predicted class label of gender with true data (i.e., manually checked data)

We also note, however, that although individual users are less likely to be misclassified by manual checks than by trained algorithms, it is still possible that we introduced bias into the process. For instance, we may have incorrectly identified the age categories of users who looked young or old to a human observer, or of users who posted outdated photos. Moreover, we cannot rule out the possibility that certain self-presentation behaviours (For instance, use of multiple languages, or frequent changes of home locations) were tied to the migration event, imposing certain limitations. Hence, it is possible that in certain circumstances, an automatic evaluation would have performed better than we would otherwise expect.

Table 2: Average precision, recall and F1 scores for age comparing our predicted class label of gender with true data (i.e., manually checked data)

	precision	recall	f1-score	support
<=18	0.41	0.81	0.54	16
19-29	0.89	0.83	0.86	251
30-39	0.88	0.81	0.84	182
>=40	0.79	0.88	0.83	106
accuracy	0.83	0.83	0.83	0.83
macro avg	0.74	0.83	0.77	555
weighted avg	0.85	0.83	0.84	555

3.4 Detecting new social connections

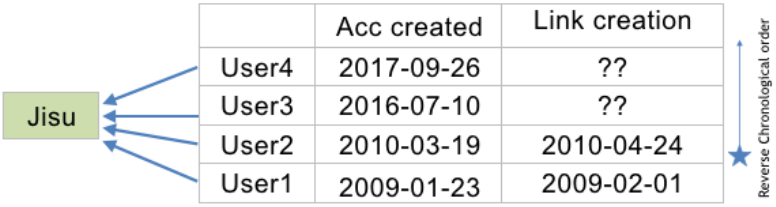
To study how the composition of a migrant's social network changed before and after migration, we need to understand when the social links were created. To do so, we applied the method developed by Meeder et al. [2011]. In their study, they inferred the date of link creation based on two pieces of information: first, the date when the account was created; and, second, the list of friends[8]/followers[9], which the Twitter API[10] provides in the reverse chronological order of link creation between two users[11], as shown in Figure 3. Based on this information, the date of link creation for an alter user u with the ego user v would be the maximum of the account creation date of all the users who followed the ego user v before the user u. This method was shown to have a high accuracy level, with the typical error for inferring the date and the time of link creation being up to a few minutes only. It should, however, be noted that this analysis was conducted using data from celebrities who had large networks with many followers and friends, and that larger networks can provide tighter temporal bounds. By contrast, the users in our data did not have as many friends or followers as these celebrities. For this reason, we expect the precision of our estimates to be

[8] "*friends* refers as the Twitter users that a specific user follows (e.g., following)." as defined by Twitter (https://developer.twitter.com/en/docs/twitter-api/v1/accounts-and-users/follow- search-get-users/overview).

[9] "*Followers* refers tc

[10] Application Progra

[11] https://developer.tv rs/api-reference/get-followers-list



	Acc created	Link creation
User4	2017-09-26	??
User3	2016-07-10	??
User2	2010-03-19	2010-04-24
User1	2009-01-23	2009-02-01

Figure 3: Example of friendship network data on Twitter: We have users 1 to 4 (alter users u) following Jisu (ego user v) and the dates of when these users created their accounts. The link creation column shows an illustrative example of the dates when the user 1 and 2 started following Jisu.

lower. Nevertheless, as we do not require exact time up to the second of the link creation date, this may not be an issue here.

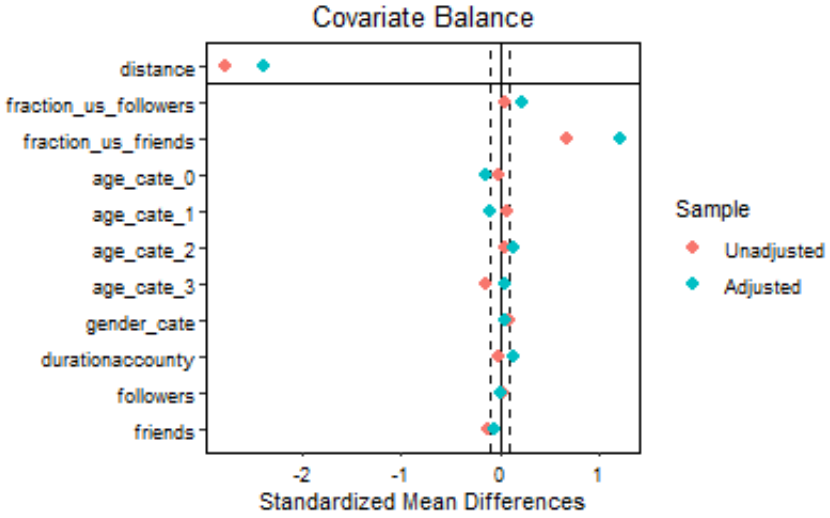


Figure 4: Histograms showing the balance for categorical variables before (‘unadjusted’) and after (‘adjusted’) matching where *fraction_us_friends* and *fraction_us_followers* mean fraction of U.S. based friends and followers, *age_cate_0* means age below 18, *age_cate_1* means age between 19-29, *age_cate_2* means age between 30-39 and *age_cate_3* means age above 40. The *durationaccounty* signifies account age.

During this process, we also collected information on the locations of the users’ friends and followers at a country level. To obtain their locations, we relied on the self-reported information on the users’ Twitter profiles instead of on geo-tagged tweets. We did so because we assume that most users state their usual location on their profile, and do not change it when they go on short trips or vacations. Nevertheless, the profile location information is not fully reliable, as it is self-reported. Users are, for instance, free to state their country of residence, city, or even coordinates. To obtain uniform location information at a country level, we used a dictionary of country names taken from http://www.geonames.org/. This enabled us to obtain information on how many friends and followers were added before and after migration, and where these friends and followers were located.

3.5 Matching migrants and non-migrants

In this project, we prioritise having reliable and high-quality information on users over the size of the data set. Hence, after obtaining complete data for both migrants and non-migrants, we performed 1:1 propensity score matching using the logistic regression model [Stuart et al., 2011] to pair migrants with non-migrants who had similar levels of activity on Twitter and similar characteristics. Comparing the friendship patterns among this placebo group with those among migrants over time (in particular, the number of links to other users in the U.S.) enabled us to isolate the effect of the migration event. The application of this matching technique was particularly important in the context of the United States, as it is the country with the largest Twitter user population. Because of this large user reservoir, there was a certain baseline rate for the creation of U.S.-based friend/follower links, irrespective of migration. Comparing similar migrants and non-migrants helped to account for this baseline rate. As was mentioned previously, we have chosen to focus on migrants in the United States. Our data include about 1300 migrants individuals who migrated to the U.S. Therefore, we performed the matching technique to select non-migrants (i.e., individuals who never changed their usual country of residence) whose characteristics were similar to those of the migrants in the U.S. Here, the degree of similarity was measured based on the users' characteristics that we could observe from Twitter (i.e., age, gender, age of Twitter account, number of followers, friends, country of origin) and on the fraction of the users' friends and followers who were based in the U.S.

Through this process, we were able to match up 1163 migrants and 1163 non-migrants. In Figure 4, we can see that the covariates were more balanced after matching. The "distance" in the first row of Figure 4 is the overall difference in propensity scores. We can also see that the overall distance was reduced after matching, which suggests that we were able to pair migrants and non-migrants with similar characteristics, and that the overall balance improved after matching. It should, however, be noted that the balance improved in most of the covariates except for the fraction of U.S.-based followers and friends and first two age categories. This may be because there were large differences in these variables between migrants and non-migrants from the beginning.

4. Results

We begin this section by comparing the distributions of the fraction of followers and friends for both migrants and non-migrants. Here, we compare the distribution of the fraction of followers (or friends) instead of the absolute number of followers (or friends), because it can provide a clearer picture of how the structure of the network changed after migration. By considering the total number of new followers (or friends) added before and after migration regardless of the user's location, we are able to observe whether an immigrant was indeed adding new social links in the United States, or whether only a small proportion of the user's new social connections were based in the U.S.

Moreover, we compare the distribution of the followers and the friends of migrants and non-migrants. Using the inferred migration date of the matched migrants, we assign a "placebo" date of migration to non-migrants to enable us to compare the shift in the distribution of the fraction of friends for both migrants and non-migrants. This allows us to purely observe the effects of the migration event. If we observe an equal shift in the distribution of the fraction of the user's U.S.-based friends before and after the placebo date of migration, this would indicate that the migration event did not play a role in the expansion of the Twitter user's social network in the United States.

In Figure 5, we observe a shift towards the right in the fraction of U.S. based-followers for both migrants and non-migrants. For migrants, the average of this fraction increased from 0.14 to 0.165. Using the placebo migration date that we assigned to non-migrants, we also show for non-migrants the distribution of the fraction of followers residing in the U.S. before and after the placebo migration date. Looking at the right panel of Figure 5, we can see that this distribution increased from 0.036 to 0.042. The results of the t-tests for both groups indicate that the means were different before and after migration (*t*-statistic value of -4.14 for migrants and -2.63 for non-migrants with a *p*-value of 3.55×10^{-5} and 0.008, respectively). Thus, we observe that the percentage change for both groups was approximately the same. Our finding that there was an equal shift in the distribution suggests that the migration event did not necessarily play a role in the expansion of the Twitter users' follower networks in the United States.

When we compare the distribution of the fraction of friends residing in the United States before and after migration for migrants (left panel) and for non-migrants (right panel) in Figure 6, we see a different pattern. In the left panel, we observe a shift in the distribution towards the right, indicating that the fraction of friends residing in the U.S. increased after migration for migrants. To be more specific, the mean value of the fraction of friends who were based in the U.S. increased, on average, from 0.214 before migration to 0.25 after migration. We further tested whether the difference between the means of two groups was significant. An independent two-sided t-test showed that the means were significantly different from each other (*t*-statistic value of -4.35 and a *p*-value of 1.42×10^{-5}).

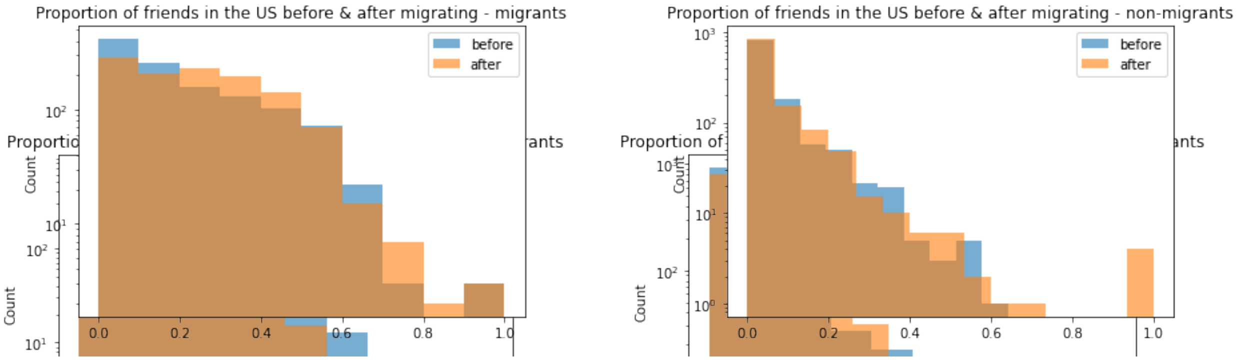


Figure 6: Distribution of fraction of friends residing in the United States before and after migration for migrants (left) and for propensity-score matched non-migrants (right).

Figure 5: Distribution of fraction of followers residing in the United States before and after migration for migrants (left) and for propensity score-matched non- migrants (right).

For non-migrants, we look at the fraction of their friends who were based in the U.S. before and after the *placebo* date of migration. As the right panel of Figure 6 illustrates, we see no shift in the distribution after migration for non-migrants, meaning that no new U.S.-based friends were added. Before the *placebo* date of migration, about 6.1% of the migrants' friends were from the U.S. This share increased only slightly after the *placebo* date of migration, to 6.2%. However, the results of the t-test showed that the average fractions of U.S.-based friends before and after migration were not statistically different from each other (*t*-statistic value of -0.2 with a *p*-value of 0.84). This indicates that there is not enough evidence to conclude that the two distributions had different means. As we observe a clear shift in the distribution for migrants but not for non-migrants, our assumption that the migration event played a significant role in expanding Twitter users' friends' networks in the United States is confirmed.

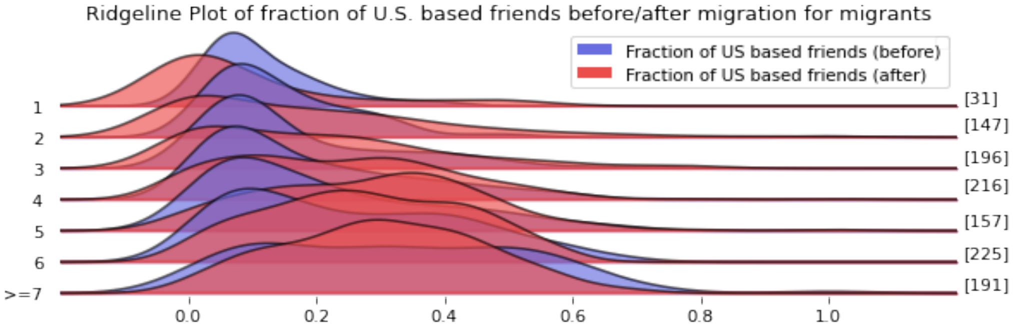


Figure 7: Ridgeline plot of fraction of U.S.-based friends after migration for migrants. The y-axis represents number of years spent in the U.S. The numbers in squared bracket indicate the number of populations considered for each number of years since migration.

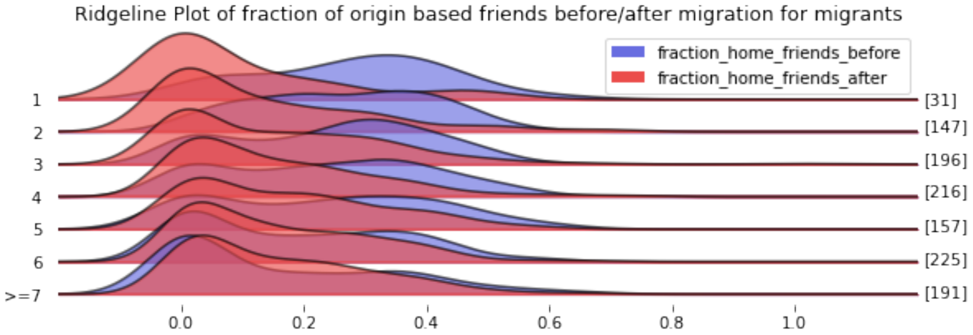


Figure 8: Ridgeline plot of fraction of origin-based friends after migration for migrants. The y-axis represents number of years spent in the U.S. The numbers in squared bracket indicate the number of populations considered for each number of years since migration.

To further improve our understanding of when the migrants began to have more U.S.-based friends on Twitter, we split the distributions by how many years migrants have been residing in the U.S. In other words, we examine how the fraction of friends migrants had before and after migration differed for those, for instance, who have been residing in the U.S. for two years. In Figure 7, we observe that as the number of years migrants spent in the U.S. increased, the fraction of friends they had after migration shifted towards the right. This indicates that migrants who had been in the U.S. for longer periods of time had more U.S.- based friends on Twitter. More specifically, we observe a clear shift for migrants starting in the third year since migration. On the other hand, we also observe that the distributions of the fraction of friends before and after migration were identical in the first year since migration and in the seventh year since migration or later. The results of the t-tests also indicate that the average fraction of U.S.-based friends before and after migration were identical for these years (*t*-statistic value of -0.8 and *p*-value of 0.43 for the first year and *t*-statistic value of -1 and *p*-value of 0.31 for the seventh year or later). These results suggest that the immigrants needed some time to expand their friendship networks on Twitter after migration. Interestingly, immigrants who had been in the U.S. for more than seven years had a relatively large fraction of U.S.-based friends before migrating, but also added a larger additional fraction of friends after migrating. Note that here we are observing different sets of users as the number of years spent in the U.S. increased (the number of observed populations for each year is indicated in squared bracket), meaning that different set of users are contributing to the changes that we observe each year.

We further test the robustness of the ridgeline plot of the fraction of U.S.-based friends after migration. We do this to check whether the changes in the fraction of U.S.-based friends were caused by changes in the composition of underlying population. To address this issue, we redo plot 7 with the calendar year of migration, instead of for the years for which we have a comparable number of users. We find that the number of users was similar (about 200 users) for the calendar years 2013 to 2016. Hence, we report the ridgeline plot for these four years. In Figure 10 in the appendix, we observe somewhat similar patterns across the years. In particular, we see that the users had, on average, a higher fraction of U.S.-based friends after migration than before migration. This tells us that the period effect on overall changes in the fraction of U.S.-based friends before and after migration event was small.

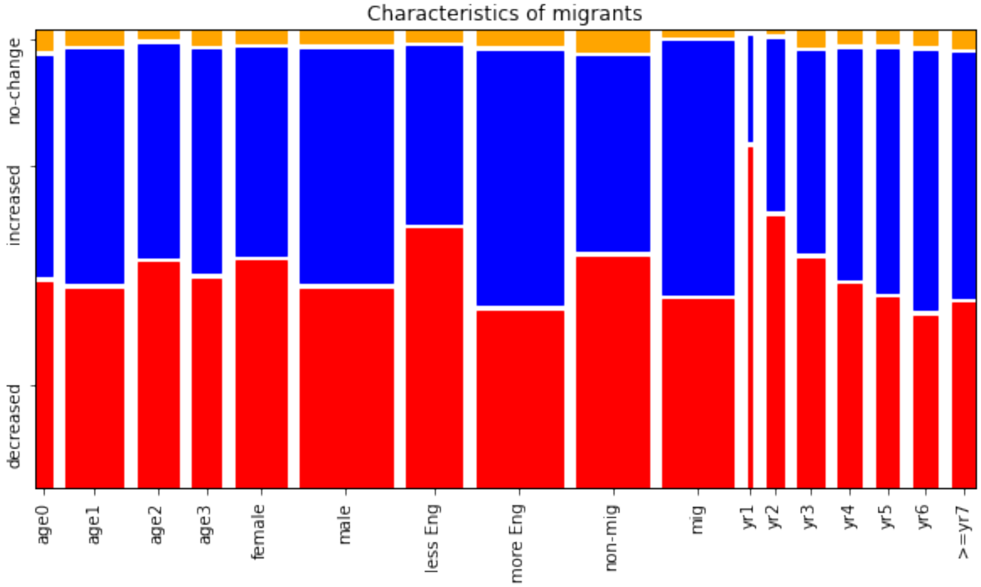


Figure 9: Correspondence analysis mosaic plot of Twitter characteristics of users: x-axis shows different characteristics of users and y-axis shows whether these characteristics are associated with either decrease, increase or no-changes in the proportion of U.S.-based friends. *age* indicates age groups from 0 to 3*, more/less Eng* indicates whether a user has tweeted more or less in English after migration, *mig* and *non-mig* indicate migrants, and non-migrants, *yr* indicates years since migration from one to seven and above

We also examine how the migrants' connections to their origin country changed before and after migration. In Figure 8, we show the same distributions, but for the fraction of origin-based friends of migrants. We see that the pattern differs from that in the previous figure, as the fraction of origin-based friends after migration stayed relatively small throughout the analysed period. This tells us that migrants did not add as many origin-based friends after migration as they did before migration. Note, however, that we observe a different pattern for migrants in the seventh year since migration or later: i.e., they had similar fractions of origin-based friends both before and after migration. The t-test results also tell us that the means were not significantly different from each other (*t*-statistic value of 0.076 and a *p*-value of 0.94). Interestingly, as more time since migration passed, the migrants also tended to also add more origin-based friends on Twitter.

As we have observed a clear shift in the proportion of the migrants' friends who were based in the U.S. after migration, we further extend the analysis to study the characteristics of these users based on several features that we can observe from the data. To do so, we first divide the users according to their fraction of U.S.-based friends into three groups: those who had a larger fraction (called "more friends"), a smaller fraction (called "less friends"), and no change (called "no change") in the fraction of U.S.-based friends after migration. We then look at the frequency of the following categorical variables: whether a user was or was not a migrant (named "mig" and "non-mig"), age category, gender, whether more English tweets were posted after migration (named "more Eng" and "less Eng"), and the number of years since migration. With this contingency table, we perform a correspondence analysis that allows us to study the dependencies between our variable of interest, i.e., the fraction of U.S.-based friends in the three groups, and other categorical variables.

First, the Chi-square test tells us that the variables have a statistically significant association (Chi-square value of 219.19 and *p*-value of 0.0004). To help us understand these associations in more detail, Figure 9 illustrates the relationships between these categories. Here, we observe that having more friends on Twitter after migration is most closely associated with tweeting more in English after migration; being male; being between the ages of 19 and 29 (*age1* from the figure); having a varying number of years since migration, including being in the fourth, fifth, sixth, or seventh year or later since migration; and being a migrant. By contrast, we find that having fewer friends on Twitter after migration is most closely associated with tweeting less in English after migration; being between the ages of 30 and 39 or being age 40 or older (*age2* and *age3* from the figure); being female; being in the first, second, or third year since migration; and being a non-migrant. Note, however, that the third year since migration is the period when we begin to observe a shift in the fraction of U.S.-based friends before and after migration as is shown in Figure 7. Moreover, for migrants who migrated for more than seven years ago, we observe no significant difference in the two distributions.

5. Discussion

In this study, through a careful design of the data collection process and various analyses, we sought to identify migrant users of Twitter, and to understand the evolution of their new social connections on the platform after migration. Thanks to the availability of these data, we were able to observe that the fraction of a migrant's friends who were based in the destination country increased, on average, from 0.21 to 0.25 after migration. Our comparison of the behaviour of matched non-migrants and migrants suggest that this increase was indeed due to the migration event. To be more specific, we found that the longer a migrant stayed in the U.S., the more U.S.-based friends they had on Twitter. Interestingly, however, we observed that a clear shift in the fraction of friends did not start until the third year since migration. Further analysis indicated that this increase in the fraction of friends was associated with several user characteristics. For instance, we observed that being a male migrant between ages 19 and 29 who tweeted more in English after migration, having migrated four or more years ago, and being a migrant were most closely associated with an increase in the fraction of friends in the destination country. Additionally, when we looked at the fraction of origin-based friends in general, we noticed that migrants did not add as many origin-based friends after migration as they did before migration. Nonetheless, after the migrants had been living in the destination country for a longer period of time, they again tended to add more origin-based friends on Twitter. However, unlike the patterns we observed for friend connections, we found no evidence that the migration event was associated with an expansion of the Twitter users' follower networks in the destination country.

Thanks to the availability of Twitter data, we were able to capture the process of social integration among users of this specific platform. We were able to obtain data that contained detailed information on migrants' (online) social networks, including precise information on how many friends were added to their networks before and after migration. These kinds of data are not available from surveys. Nevertheless, Twitter data have some limitations. First, it is important to mention that as we prioritised having reliable and high-quality data on users, we obtained a relatively small data set that does not allow us to generalise our findings. Additional experiments with Twitter users should be done in order to provide further validation of our results for this platform. Moreover, there are many challenges that arise when seeking to identify migrants from geo-located tweets. Thus, it is possible that we identified members of a highly mobile population, such as short-term visitors (e.g., students) and business and leisure travellers [Armstrong et al., 2021], rather than migrants. On the related note, we relied here on the methodology developed by Wang et al. [2019] to obtain demographic information. As was mentioned previously, although the algorithm had a high accuracy level, it did not allow us to break down age categories into finer classes. Furthermore, the binary gender category we used does not reflect everyone's reality.

Second, people's social networks on Twitter differ from their real-world social networks, and from the network typologies of other social media platforms. On Twitter, individual users are free to follow other users as long as the accounts are publicly open. Hence, users can also interact with other users they do not know personally. This is different from other social media platforms such as Facebook, where the interactions are between close and mutual friends. To tackle this issue, we plan to further investigate how much of this social integration on Twitter is reflected in real life. Xie et al. [2012] examined the extent to which users' Twitter networks are reflected in their offline, real-life social networks. We also intend to explore this issue in order to better understand the magnitude of this social network effect in real life. Furthermore, based on the users' self-reported information, we were able to infer the current locations of their friends. However, this did not provide us with any information on whether these friends were or were not migrants themselves. In the future, we plan to also investigate the countries of origin of immigrants' friends in order to better understand whether they are integrating into the host society, or are remaining within their community of origin. This is also a crucial factor to consider when studying social integration, as the failure of immigrants to integrate into the local population could lead to social division between different communities. Without a doubt, the analysis of new social links that this paper has presented does not reflect all aspects of social integration. Nevertheless, we believe that our work shows how offline (re)location plays a significant role in the formation of immigrants' online networks.

Finally, and importantly, we want to emphasise that no personal information has been published at any stage of the research. All of the results are aggregated at the country level to mask any factors that can be traced back to a specific Twitter user.

References

1. Caitrin Armstrong, Ate Poorthuis, Matthew Zook, Derek Ruths, and Thomas Soehl. Challenges when identifying migration from geo-located twitter data. EPJ Data Science, 10(1):1, 2021.
2. Dany Bahar, Hillel Rapoport, and Riccardo Turati. Birthplace diversity and eco- nomic complexity: Cross-country evidence. Research Policy, page 103991, 2020.
3. Pierre Bourdieu. The forms of capital. handbook of theory and research for the sociology of education. jg richardson. New York, Greenwood, 241(258):19, 1986.
4. Guanghua Chi, Joshua E Blumenstock, Lada Adamic, et al. Who ties the world together? evidence from a large online social network. In International Confer- ence on Complex Networks and Their Applications, pages 451–465. Springer, 2019.
5. Frédéric Docquier and Hillel Rapoport. Globalization, brain drain, and development. Journal of economic literature, 50(3):681–730, 2012.
6. Antoine Dubois, Emilio Zagheni, Kiran Garimella, and Ingmar Weber. Studying migrant assimilation through Facebook interests. In International Conference on Social Informatics, pages 51–60. Springer, 2018.
7. Birgit Glorius, Stefan Kordel, Tobias Weidinger, Miriam Bu¨rer, Hanne Schneider, and David Spenger. Is social contact with the resident population a prerequisite of well-being and place attachment? the case of refugees in rural regions of germany. Frontiers in Sociology, 5:114, 2020.

8. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

9. Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE con- ference on computer vision and pattern recognition, pages 4700–4708, 2017.

10.Ade Kearns and Elise Whitley. Getting there? the effects of functional factors, time and place on the social integration of migrants. Journal of ethnic and migration studies, 41(13):2105–2129, 2015.

11. Jisu Kim, Alina Sîbu, Fosca Giannotti, and Giulio Rossetti. Characterising different communities of twitter users: Migrants and natives. In Conference proceedings COMPLEX NETWORKS 2021, page to appear, 2021a.

12.Jisu Kim, Alina Sîbu, Giulio Rossetti, Fosca Giannotti, and Hillel Rapoport. Origin and destination attachment: study of cultural integration on twitter. arXiv preprint arXiv:2102.11398, -forthcoming in EPJ Data Science, 2022b.

13. Vili Lehdonvirta and Pekka Räsänen. How do young people identify with online and offline peer groups? a comparison between uk, spain and japan. Journal of Youth Studies, 14(1):91–108, 2011.

14. Mattia Mazzoli, Boris Diechtiareff, Antònia Tugores, Willian Wives, Natalia Adler, Pere Colet, and José J Ramasco. Migrant mobility flows characterized with digital data. PloS one, 15(3):e0230264, 2020.

15. Brendan Meeder, Brian Karrer, Amin Sayedi, R Ravi, Christian Borgs, and Jen- nifer Chayes. We know who you followed last summer: inferring social link creation times in twitter. In Proceedings of the 20th international conference on World wide web, pages 517–526, 2011.

16. Alejandro Portes. Social capital: Its origins and applications in modern sociology. Annual review of sociology, 24(1):1–24, 1998.

17. Robert D. Putnam. Bowling alone: America’s declining social capital. Journal of Democracy, 6(1):65–78, 1995. URL http://www.journalofdemocracy.org/.

18. James E Rauch. Business and social networks in international trade. Journal of economic literature, 39(4):1177–1203, 2001.

19. James E Rauch and Vitor Trindade. Ethnic chinese networks in international trade. Review of Economics and Statistics, 84(1):116–130, 2002.

20. Maurice Schiff. Social capital, labor mobility, and welfare: The impact of uniting states. Rationality and Society, 4(2):157–175, 1992.

21. Alina Sîrbu, Gennady Andrienko, Natalia Andrienko, Chiara Boldrini, Marco Conti, Fosca Giannotti, Riccardo Guidotti, Simone Bertoli, Jisu Kim, Cristina Ioana Muntean, et al. Human migration: the big data perspective. International Journal of Data Science and Analytics, pages 1–20, 2020.

22. Ian Stewart, René D Flores, Timothy Riffe, Ingmar Weber, and Emilio Zagheni. Rock, Rap, or Reggaeton?: Assessing Mexican Immigrants' Cultural Assimila- tion Using Facebook Data. In The World Wide Web Conference, pages 3258– 3264. ACM, 2019.

23. Elizabeth A Stuart, Gary King, Kosuke Imai, and Daniel Ho. Matchit: nonpara- metric preprocessing for parametric causal inference. Journal of statistical soft- ware, 2011.

24. Eric M Uslaner. Volunteering and social capital: how trust and religion shape civic participation in the United States. Routledge, 2003.

25. Loïc JD Wacquant and Pierre Bourdieu. An invitation to reflexive sociology. Polity Cambridge, 1992.

26. Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flo¨ck, and David Jurgens. Demographic infer- ence and representative population estimates from multilingual social media data. In The World Wide Web Conference, WWW '19, page 2056–2067, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313684. URL https://doi.org/10.1145/3308558.3313684.

27. Susanne Wessendorf and Jenny Phillimore. New migrants' social integration, embedding and emplacement in superdiverse contexts. Sociology, 53(1):123– 138, 2019.

28. Wei Xie, Cheng Li, Feida Zhu, Ee-Peng Lim, and Xueqing Gong. When a friend in twitter is a friend in life. In Proceedings of the 4th Annual ACM Web Science Conference, pages 344–347, 2012.

29. In-Jin Yoon. Social integration and well-being of north korean migrants in south korea. Journal of Social Issues, 2021.

Appendix

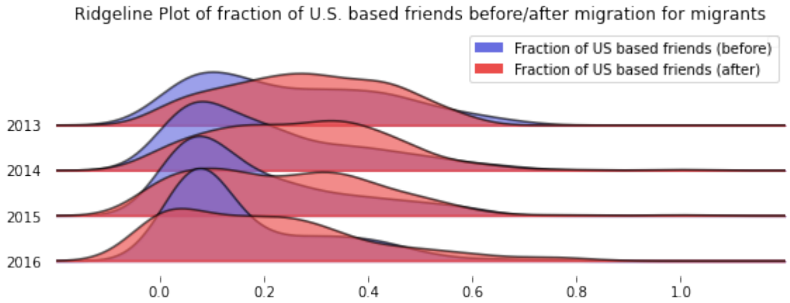


Figure 10. Ridgeline plot of fraction of U.S.-based friends before and after migration for migrants per calendar year of migration event.

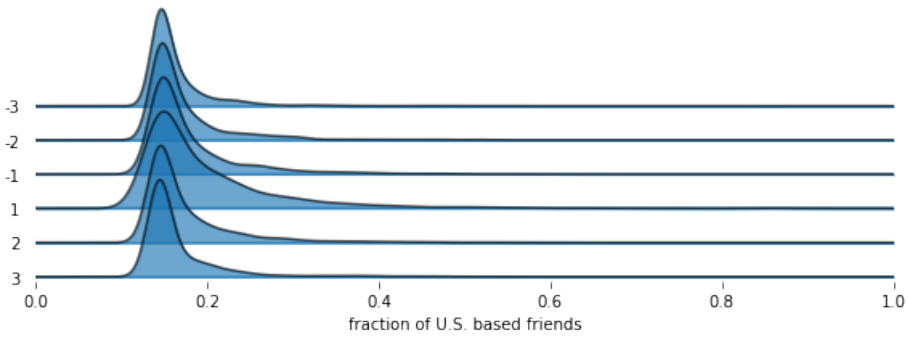


Figure 11: Ridgeline plot of fraction of U.S. based friends before and after migration for migrants each year. The number of users is the same at each y-axis as we are observing how many U.S. based friends have an immigrant added each year. E.g., 1 on the y-axis means first year of migration, and -1 means a year before migration. We observe that immigrants on Twitter have added most of the new U.S. based friends in the first year after migration with average of 0.077. Then we observe a decline in the fraction of U.S. based friends being added from second year and above (average of 0.045 in the second year and 0.033 in the third year). Furthermore, immigrants tend to also add some fraction of U.S. based friends prior to migration. More precisely, we observe that immigrants added, on average, 5.2% of new U.S. based friends a year prior to migration, 4.2% in the two years prior to migration and 3% in the three years prior to migration.